

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

RECEIVED  
CENTRAL FAX CENTER

SEP 18 2006

**Amendments to the Claims**

This listing of claims will replace all prior versions, and listings, of claims in the application:

**Listing of Claims:**

- 1           1.       (currently amended): A system for grouping clusters of  
2       semantically scored documents electronically stored in a data corpus, comprising:  
3           a scoring module determining a score, which is assigned to at least one  
4       concept that has been extracted from a plurality of electronically-stored  
5       documents, wherein the score is based on at least one of a frequency of  
6       occurrence of the at least one concept within at least one such document, a  
7       concept weight, a structural weight, and a corpus weight; [[and]]  
8           a clustering module forming clusters of the documents by applying  
9       evaluating the score for the at least one concept [[to]] of each document for a best  
10      fit criterion for each such document to the clusters and assigning each document  
11      to the cluster with the best fit; and  
12           a threshold module dynamically determining a threshold for each cluster  
13      based on similarities between the documents grouped into the cluster and a center  
14      of the cluster, and reassigning those documents having similarities outside the  
15      threshold.
- 1           2.       (original): A system according to Claim 1, further comprising:  
2           the scoring module calculating the score as a function of a summation of  
3           at least one of the frequency of occurrence, the concept weight, the structural  
4           weight, and the corpus weight of the at least one concept.
- 1           3.       (original): A system according to Claim 2, further comprising:  
2           a compression module compressing the score through logarithmic  
3           compression.
- 1           4.       (original): A system according to Claim 1, further comprising:

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

2 the scoring module calculating the concept weight as a function of a  
3 number of terms comprising the at least one concept.

1 5. (original): A system according to Claim 1, further comprising:  
2 the scoring module calculating the structural weight as a function of a  
3 location of the at least one concept within the at least one such document.

1 6. (original): A system according to Claim 1, further comprising:  
2 the scoring module calculating the corpus weight as a function of a  
3 reference count of the at least one concept over the plurality of documents.

1 7. (original): A system according to Claim 1, further comprising:  
2 the scoring module forming the score assigned to the at least one concept  
3 to a normalized score vector for each such document, determining a similarity  
4 between the normalized score vector for each such document as an inner product  
5 of each normalized score vector, and applying the similarity to the best fit  
6 criterion.

1 8. (original): A system according to Claim 1, further comprising:  
2 the clustering module evaluating a set of candidate seed documents  
3 selected from the plurality of documents, identifying a set of seed documents by  
4 applying the score for the at least one concept to a best fit criterion for each such  
5 candidate seed document, and basing the best fit criterion on the score of each  
6 such seed document.

1 9. (currently amended): A method for grouping clusters of  
2 semantically scored documents electronically stored in a data corpus, comprising:  
3 determining a score, which is assigned to at least one concept that has  
4 been extracted from a plurality of electronically-stored documents, wherein the  
5 score is based on at least one of a frequency of occurrence of the at least one  
6 concept within at least one such document, a concept weight, a structural weight,  
7 and a corpus weight; [[and]]

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

8        forming logically-grouped clusters of the documents by applying  
9        evaluating the score for the at least one concept [[to]] of each document for a best  
10       fit criterion for each such document: to the clusters and assigning each document  
11       to the cluster with the best fit;  
12       dynamically determining a threshold for each cluster based on similarities  
13       between the documents grouped into the cluster and a center of the cluster; and  
14       reassigning those documents having similarities outside the threshold.

1       10.    (original): A method according to Claim 9, further comprising:  
2       calculating the score as a function of a summation of at least one of the  
3       frequency of occurrence, the concept weight, the structural weight, and the corpus  
4       weight of the at least one concept.

1       11.    (original): A method according to Claim 10, further comprising:  
2       compressing the score through logarithmic compression.

1       12.    (original): A method according to Claim 9, further comprising:  
2       calculating the concept weight as a function of a number of terms  
3       comprising the at least one concept.

1       13.    (original): A method according to Claim 9, further comprising:  
2       calculating the structural weight as a function of a location of the at least  
3       one concept within the at least one such document.

1       14.    (original): A method according to Claim 9, further comprising:  
2       calculating the corpus weight as a function of a reference count of the at  
3       least one concept over the plurality of documents.

1       15.    (original): A method according to Claim 9, further comprising:  
2       forming the score assigned to the at least one concept to a normalized  
3       score vector for each such document;  
4       determining a similarity between the normalized score vector for each  
5       such document as an inner product of each normalized score vector; and

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

6 applying the similarity to the best fit criterion.

1 16. (original): A method according to Claim 9, further comprising:  
2 evaluating a set of candidate seed documents selected from the plurality of  
3 documents;  
4 identifying a set of seed documents by applying the score for the at least  
5 one concept to a best fit criterion for each such candidate seed document; and  
6 basing the best fit criterion on the score of each such seed document.

1 17. (currently amended): A computer-readable storage medium  
2 holding code for ~~performing the method of Claim 9, grouping clusters of~~  
3 semantically scored documents electronically stored in a data corpus, comprising:  
4 code for determining a score, which is assigned to at least one concept that  
5 has been extracted from a plurality of electronically-stored documents, wherein  
6 the score is based on at least one of a frequency of occurrence of the at least one  
7 concept within at least one such document, a concept weight, a structural weight,  
8 and a corpus weight;  
9 code for forming logically-grouped clusters of the documents by  
10 evaluating the score for the at least one concept of each document for a best fit to  
11 the clusters and assigning each document to the cluster with the best fit;  
12 code for dynamically determining a threshold for each cluster based on  
13 similarities between the documents grouped into the cluster and a center of the  
14 cluster; and  
15 code for reassigning those documents having similarities outside the  
16 threshold.

1 18. (currently amended): A system for providing efficient document  
2 scoring of concepts within [[a]] and clustering of documents in an electronically-  
3 stored document set, comprising:  
4 a scoring module scoring a document in an electronically-stored document  
5 set, comprising:

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

6 a frequency module determining a frequency of occurrence of at  
7 least one concept within a ~~document retrieved from the document set; and~~  
8 document;  
9 a concept weight module analyzing a concept weight reflecting a  
10 specificity of meaning for the at least one concept within the document;  
11 a structural weight module analyzing a structural weight reflecting  
12 a degree of significance based on structural location within the document for the  
13 at least one concept;  
14 a corpus weight module analyzing a corpus weight inversely  
15 weighing a reference count of occurrences for the at least one concept within the  
16 document; and  
17 a scoring evaluation module evaluating a score to be associated  
18 with the at least one concept as a function of the frequency, concept weight,  
19 structural weight, and corpus weight. weight; and  
20 a clustering module grouping the documents by score into a plurality of  
21 clusters, comprising:  
22 a cluster seed module identifying candidate seed documents, which  
23 are each assigned as a seed document into a cluster with a center most similar to  
24 the seed document, and assigning each non-seed document to the cluster with the  
25 best fit; and  
26 a threshold module dynamically determining a threshold for each  
27 cluster based on similarities between the documents in each cluster and the cluster  
28 center, and reassigning the documents with similarities outside the threshold.

1 19. (currently amended): A system according to Claim 18, further  
2 comprising:  
3 the scoring module evaluating the score ~~substantially~~ in accordance with  
4 the formula:

$$5 \quad S_i = \sum_{j=1}^n f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

6 where  $S_i$  comprises the score,  $f_{ij}$  comprises the frequency,  $0 < cw_{ij} \leq 1$  comprises  
7 the concept weight,  $0 < sw_{ij} \leq 1$  comprises the structural weight, and  $0 < rw_{ij} \leq 1$   
8 comprises the corpus weight for occurrence  $j$  of concept  $i$ .

1 20. (currently amended): A system according to Claim 19, further  
2 comprising:  
3 the concept weight module evaluating the concept weight substantially in  
4 accordance with the formula:

$$5 \quad cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

6 where  $cw_{ij}$  comprises the concept weight and  $t_{ij}$  comprises a number of terms for  
7 occurrence  $j$  of each such concept  $i$ .

1 21. (currently amended): A system according to Claim 19, further  
2 comprising:  
3 the structural weight module evaluating the structural weight substantially  
4 in accordance with the formula:

$$5 \quad sw_{ij} = \begin{cases} 1.0, & \text{if } (j \approx \text{SUBJECT}) \\ 0.8, & \text{if } (j \approx \text{HEADING}) \\ 0.7, & \text{if } (j \approx \text{SUMMARY}) \\ 0.5 & \text{if } (j \approx \text{BODY}) \\ 0.1 & \text{if } (j \approx \text{SIGNATURE}) \end{cases}$$

6 where  $sw_{ij}$  comprises the structural weight for occurrence  $j$  of each such concept  $i$ .

1 22. (currently amended): A system according to Claim 19, further  
2 comprising:  
3 the corpus weight module evaluating the corpus weight substantially in  
4 accordance with the formula:

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

$$rw_{ij} = \begin{cases} \left( \frac{T - r_{ij}}{T} \right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

where  $rw_{ij}$  comprises the corpus weight,  $r_{ij}$  comprises a reference count for occurrence  $j$  of each such concept  $i$ ,  $T$  comprises a total number of reference counts of documents in the document set, and  $M$  comprises a maximum reference count of documents in the document set.

23. (currently amended): A system according to Claim 19, further comprising:  
a compression module compressing the score ~~substantially~~ in accordance with the formula:

$$S'_i = \log(S_i + 1)$$

where  $S'_i$  comprises the compressed score for each such concept  $i$ .

24. (original): A system according to Claim 18, further comprising:  
a global stop concept vector cache maintaining concepts and terms; and  
a filtering module filtering selection of the at least one concept based on the concepts and terms maintained in the global stop concept vector cache.

25. (original): A system according to Claim 18, further comprising:  
a parsing module identifying terms within at least one document in the document set, and combining the identified terms into one or more of the concepts.

26. (original): A system according to Claim 25, further comprising:  
the parsing module structuring each such identified term in the one or more concepts into canonical concepts comprising at least one of word root, character case, and word ordering.

27. (original): A system according to Claim 25, wherein at least one of nouns, proper nouns and adjectives are included as terms.

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

1           28.     (original): A system according to Claim 18, further comprising:  
2           a plurality of candidate seed documents;  
3           a similarity module determining a similarity between each pair of a  
4 candidate seed document and a cluster center;  
5           a clustering module designating each such candidate seed document  
6 separated from substantially all cluster centers with such similarity being  
7 sufficiently distinct as a seed document, and grouping each such candidate seed  
8 document not being sufficiently distinct into a cluster with a nearest cluster  
9 center.

1           29.     (original): A system according to Claim 28, further comprising:  
2           a plurality of non-seed documents;  
3           the similarity module determining the similarity between each non-seed  
4 document and each cluster center; and  
5           the clustering module grouping each such non-seed document into a  
6 cluster having a best fit, subject to a minimum fit criterion.

1           30.     (original): A system according to Claim 29, further comprising:  
2           a normalized score vector for each document comprising the score  
3 associated with the at least one concept for each such concept occurring within  
4 the document; and  
5           the similarity module determining the similarity as a function of the  
6 normalized score vector associated with the at least one concept for each such  
7 document.

1           31.     (currently amended): A system according to Claim 30, further  
2 comprising:  
3           the similarity module calculating the similarity substantially in accordance  
4 with the formula:



Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

$$\cos \sigma_{AB} = \frac{\langle \bar{S}_A \cdot \bar{S}_B \rangle}{|\bar{S}_A| |\bar{S}_B|}$$

where  $\cos \sigma_{AB}$  comprises a similarity between a document  $A$  and a document  $B$ ,  
 $\bar{S}_A$  comprises a score vector for document  $A$ , and  $\bar{S}_B$  comprises a score vector for  
document  $B$ .

Claims 32-34 (canceled).

35. (currently amended): A method for providing efficient document  
scoring of concepts within [[a]] and clustering of documents in an electronically-  
stored document set, comprising:  
scoring a document in an electronically-stored document set, comprising:  
determining a frequency of occurrence of at least one concept  
within a ~~document retrieved from the document set; and~~ document;  
analyzing a concept weight reflecting a specificity of meaning for  
the at least one concept within the document;  
analyzing a structural weight reflecting a degree of significance  
based on structural location within the document for the at least one concept;  
analyzing a corpus weight inversely weighing a reference count of  
occurrences for the at least one concept within the document; and  
evaluating a score ~~to be~~ associated with the at least one concept as  
a function of the frequency, concept weight, structural weight, and corpus ~~weight-~~  
weight; and  
grouping the documents by score into a plurality of clusters, comprising:  
identifying candidate seed documents, which are each assigned as  
a seed document into a cluster with a center most similar to the seed document;  
assigning each non-seed document to the cluster with the best fit;  
dnamically determining a threshold for each cluster based on  
similarities between the documents in each cluster and the cluster center; and  
reassigning the documents with similarities outside the threshold.

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

1           36.     (currently amended): A method according to Claim 35, further  
2     comprising:  
3           evaluating the score ~~substantially~~ in accordance with the formula:

$$4 \quad S_i = \sum_{j=1}^n f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

5     where  $S_i$  comprises the score,  $f_{ij}$  comprises the frequency,  $0 < cw_{ij} \leq 1$  comprises  
6     the concept weight,  $0 < sw_{ij} \leq 1$  comprises the structural weight, and  $0 < rw_{ij} \leq 1$   
7     comprises the corpus weight for occurrence  $j$  of concept  $i$ .

1           37.     (currently amended): A method according to Claim 36, further  
2     comprising:  
3           evaluating the concept weight ~~substantially~~ in accordance with the  
4     formula:

$$5 \quad cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

6     where  $cw_{ij}$  comprises the concept weight and  $t_{ij}$  comprises a number of terms for  
7     occurrence  $j$  of each such concept  $i$ .

1           38.     (currently amended): A method according to Claim 36, further  
2     comprising:  
3           evaluating the structural weight ~~substantially~~ in accordance with the  
4     formula:

$$5 \quad sw_{ij} = \begin{cases} 1.0, & \text{if } (j \approx \text{SUBJECT}) \\ 0.8, & \text{if } (j \approx \text{HEADING}) \\ 0.7, & \text{if } (j \approx \text{SUMMARY}) \\ 0.5 & \text{if } (j \approx \text{BODY}) \\ 0.1 & \text{if } (j \approx \text{SIGNATURE}) \end{cases}$$

6     where  $sw_{ij}$  comprises the structural weight for occurrence  $j$  of each such concept  $i$ .

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

1           39.     (currently amended): A method according to Claim 36, further  
2 comprising:  
3           evaluating the corpus weight ~~substantially~~ in accordance with the formula:

$$4 \quad r_{w_{ij}} = \begin{cases} \left( \frac{T - r_{ij}}{T} \right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

5 where  $r_{w_{ij}}$  comprises the corpus weight,  $r_{ij}$  comprises a reference count for  
6 occurrence  $j$  of each such concept  $i$ ,  $T$  comprises a total number of reference  
7 counts of documents in the document set, and  $M$  comprises a maximum reference  
8 count of documents in the document set.

1           40.     (currently amended): A method according to Claim 36, further  
2 comprising:  
3           compressing the score ~~substantially~~ in accordance with the formula:  
4            $S'_i = \log(S_i + 1)$   
5 where  $S'_i$  comprises the compressed score for each such concept  $i$ .

1           41.     (original): A method according to Claim 35, further comprising:  
2           maintaining concepts and terms in a global stop concept vector cache; and  
3           filtering selection of the at least one concept based on the concepts and  
4 terms maintained in the global stop concept vector cache.

1           42.     (original): A method according to Claim 35, further comprising:  
2           identifying terms within at least one document in the document set; and  
3           combining the identified terms into one or more of the concepts.

1           43.     (original): A method according to Claim 42, further comprising:  
2           structuring each such identified term in the one or more concepts into  
3 canonical concepts comprising at least one of word root, character case, and word  
4 ordering.

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

1        44.    (original): A method according to Claim 42, further comprising:  
2        including as terms at least one of nouns, proper nouns and adjectives.

1        Claim 45 (canceled).

1        46.    (currently amended): A method according to Claim ~~[[45,]]~~ 35,  
2        further comprising:  
3        identifying a plurality of non-seed documents;  
4        determining the similarity between each non-seed document and each  
5        cluster center; and  
6        grouping each such non-seed document into a cluster with a best fit,  
7        subject to a minimum fit criterion.

1        47.    (original): A method according to Claim 46, further comprising:  
2        forming a normalized score vector for each document comprising the  
3        score associated with the at least one concept for each such concept occurring  
4        within the document; and  
5        determining the similarity as a function of the normalized score vector  
6        associated with the at least one concept for each such document.

1        48.    (currently amended): A method according to Claim 47, further  
2        comprising:  
3        calculating the similarity ~~substantially~~ in accordance with the formula:

$$4 \quad \cos \sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A| |\vec{S}_B|}$$

5        where  $\cos \sigma_{AB}$  comprises a similarity between a document  $A$  and a document  $B$ ,  
6         $\vec{S}_A$  comprises a score vector for document  $A$ , and  $\vec{S}_B$  comprises a score vector for  
7        document  $B$ .

1        Claims 49-51 (canceled).

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

1           52.     (currently amended): A computer-readable storage medium  
2     holding code for ~~performing the method of Claim 35, providing efficient~~  
3     document scoring of concepts within [[a]] and clustering of documents in an  
4     electronically-stored document set, comprising:  
5             code for scoring a document in an electronically-stored document set,  
6     comprising:  
7             code for determining a frequency of occurrence of at least one  
8     concept within a document;  
9             code for analyzing a concept weight reflecting a specificity of  
10    meaning for the at least one concept within the document;  
11            code for analyzing a structural weight reflecting a degree of  
12    significance based on structural location within the document for the at least one  
13    concept;  
14            code for analyzing a corpus weight inversely weighing a reference  
15    count of occurrences for the at least one concept within the document; and  
16            code for evaluating a score to be associated with the at least one  
17    concept as a function of the frequency, concept weight, structural weight, and  
18    corpus weight; and  
19            code for grouping the documents by score into a plurality of clusters,  
20    comprising:  
21            code for identifying candidate seed documents, which are each  
22    assigned as a seed document into a cluster with a center most similar to the seed  
23    document;  
24            code for assigning each non-seed document to the cluster with the  
25    best fit;  
26            code for dynamically determining a threshold for each cluster  
27    based on similarities between the documents in each cluster and the cluster center;  
28    and  
29            code for reassigning the documents with similarities outside the  
30    threshold.

Response to Supplemental Office Action  
Docket No. 013.0207.US.UTL

1           53. (currently amended): An apparatus for providing efficient  
2 document scoring of concepts within [[a]] and clustering of documents in an  
3 electronically-stored document set, comprising:  
4           means for scoring a document in an electronically-stored document set,  
5 comprising:  
6                   means for determining a frequency of occurrence of at least one  
7 concept within a ~~document retrieved from the document set; and~~ document;  
8                   means for analyzing a concept weight reflecting a specificity of  
9 meaning for the at least one concept within the document;  
10                  means for analyzing a structural weight reflecting a degree of  
11 significance based on structural location within the document for the at least one  
12 concept;  
13                  means for analyzing a corpus weight inversely weighing a  
14 reference count of occurrences for the at least one concept within the document;  
15 and  
16                  means for evaluating a score to be associated with the at least one  
17 concept as a function of the frequency, concept weight, structural weight, and  
18 ~~corpus weight; and~~  
19                  means for grouping the documents by score into a plurality of clusters,  
20 comprising:  
21                   means for identifying candidate seed documents, which are each  
22 assigned as a seed document into a cluster with a center most similar to the seed  
23 document;  
24                   means for assigning each non-seed document to the cluster with  
25 the best fit;  
26                   means for dynamically determining a threshold for each cluster  
27 based on similarities between the documents in each cluster and the cluster center;  
28 and  
29                  means for reassigning the documents with similarities outside the  
30 threshold.